AFRL-IF-RS-TR-2006-314
Final Technical Report
October 2006

MODELING, ANALYSIS, SIMULATION, AND SYNTHESIS OF BIOMOLECULAR NETWORKS

University of Pennsylvania

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. M301/00

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO FINAL REPORT

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Rome Research Site Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-IF-RS-TR-2006-314 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/

DANIEL J. BURNS
Work Unit Manager

/s/

JAMES A. COLLINS
Deputy Chief, Advanced Computing Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
OCT 2006	Final	Aug 01 – Dec 05

4. TITLE AND SUBTITLE
MODELING, ANALYSIS, SIMULATION, AND SYNTHESIS OF BIOMOLECULAR NETWORKS

5a. CONTRACT NUMBER
F30602-01-2-0563

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61101E

6. AUTHOR(S)
Harvey Ruben, Vijay Kumar, Oleg Sokolsky

5d. PROJECT NUMBER
BIOC

5e. TASK NUMBER
M3

5f. WORK UNIT NUMBER
01

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania
536 Johnson Pavilion
Philadelphia Pennsylvania 19104-6073

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFTC
525 Brooks Road
Rome New York 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-314

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA#06-728

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project under the DARPA BIOCOMP program integrated fundamental scientific investigations in the field of molecular systems biology, algorithm development for biomolecular modeling, and open source, object based software implementation. Major accomplishments were 1) experimental gene knockout strain investigations of the *V.fisheri* quorum sensing system that yielded a mathematical model of its regulatory proteins, 2) a model of stringent response in *E.coli* and *M.tuberculosis* describing the role of enzyme Rel_{Mtb}, 3) a first example of reachability analysis applied to a biomolecular system (lactose induction), 4) a model of tetracycline resistance that discriminates between two possible mechanisms for tetracycline diffusion through the cell membrane, and 5) a new method for investigating the 'producibility' of a metabolite by a network of chemical reactions from an available set of nutrients using sets of gene knockouts. Accomplishments in algorithm/implementation were 1) reachability and other metabolic analysis tools for non-linear biomolecular networks aiding construction of a hybrid systems-based abstraction, 2) a Systems Biology Markup Language compatible reachability algorithm using a piecewise multi-affine hybrid system method, and 3) a metabolic network producibility analysis algorithm for large scale metabolic networks predicting the possibility of producing a set of metabolites from a set of available nutrients, complementing biomass flux optimization.

15. SUBJECT TERMS
Biomolecular systems biology, modeling, gene knock out, quorum sensing, stringent response, Rel_{Mtb}, tetracycline resistance, producibility, multi-affine systems, biomass flux optimization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Daniel J. Burns
U	U	U	UL	18	19b. TELEPHONE NUMBER (*Include area code*)

Table of Contents

Summary

Project Goals

To integrate fundamental scientific investigation in the field of molecular systems biology, algorithm development for biomolecular modeling and software implementation.

Major accomplishments

Experimental investigation and modeling of bio-genetic systems

- The *V. fisheri* quorum sensing system mathematical model was constructed from experimental investigation.
    - gene knockout strains were constructed
    - strains were observed in continuous culture setting
    - effects of different regulatory proteins were investigated
    - mathematical model of the phenomenon was constructed
- The mathematical model of stringent response system in *E.coli* and *M.tuberculosis* was constructed.
    - the role of the newly discovered enzyme Rel_{Mtb} was investigated
    - data was collected from literature as well as in-house experiments
    - comprehensive mathematical model was created
- The lactose induction system was a first example of application of reachability analysis to a bio-molecular system.
- The study of tetracycline resistance is still ongoing work.
    - a mathematical model is constructed
    - mathematical model is used to discriminate between two possible mechanisms for tetracycline diffusion through the cell membrane
- We defined the notion of 'producibility' of a metabolite by a given network of chemical reactions.
    - based on stoichiometric network information
    - demonstrated use in identifying those small molecules that can be produced from an available set of nutrients
    - identified sets of genes whose knockout renders essential nutrients not producible
    - demonstrated overlapped with known essential genes.

Algorithm development and software implementation

- Algorithms were built for reachability analysis and tools for metabolic network analysis were developed.
    - reachability analysis methods developed
    - adapted to specific context of bio-molecular models
    - demonstrated that the efficient construction of a hybrid systems-based abstraction to non-linear biomolecular models was crucial to the applicability of our methods
- Reachability algorithm was developed and implemented in BioCharon.
    - takes as input a model defined in Systems Biology Markup Language (SBML)
    - converts it into a piecewise multi-affine hybrid system
    - performs reachability analysis

- We also developed and implemented a metabolic network analysis algorithm based on the theoretical results on producibility in metabolic networks.
    - handles large scale metabolic networks
    - predicts the possibility of producing a set of metabolites from a set of available nutrients
    - complements the more widespread approach of biomass flux optimization

Recommended Future Research Directions

- Building continuous dynamical models of biomolecular processes
- Development of reachability algorithms
- Building hybrid systems models based on experimental input

1. Accomplishments

1.1. Experimental investigation and mechanistic modeling of bio-genetic regulatory systems

1.1.1. Quorum sensing in *Vibrio fischeri*

Expression of lux in cpdP mutants of *V. fischeri*

We have constructed mutants of *V. fischeri* MJ-215 (Δ*luxI, ainS*) and MJ-211 (Δ*luxI*) that are defective in the *cpdP* gene. These mutants are unable to produce periplasmic 3':5'-cyclic nucleotide phosphodiesterase (3':5'-CNP) and therefore do not degrade exogenously added cAMP. With these mutants we tested the hypothesis that in the native *V. fischeri* background high levels of LuxR protein can activate *lux* operon expression in the absence of the acyl-HSLs, 3-oxo-hexanoyl-HSL and octanoyl-HSL. Previously, we had demonstrated that exogenous addition of cAMP to cultures of *V. fischeri* can stimulate expression from the *luxR* promoter. This effect provides a simple and direct way to manipulate levels of LuxR protein in the cell. Studies of the effect of added cAMP are made problematic, however, due to the presence in the cell's periplasm of a potent 3':5'-CNP that degrades extracellular cAMP. Thus, the presence of this enzyme together with the natural low permeability of cAMP presents difficulty in raising the cellular level of cAMP by exogenous addition. Our genetic elimination of the *cpdP* gene alleviates much of this problem. Initial analysis of these mutants indicates that exogenous addition of high levels of cAMP, to activate expression from the *luxR* promoter and thereby elevate cellular levels of LuxR, stimulates *lux* operon expression 3-10 fold. If supported by additional experimentation, these initial results would confirm our hypothesis that LuxR protein, at high levels, can activate *lux* operon expression in the absence of acyl-HSLs. This regulatory relationship adds an important new factor to understanding the global cellular control of luminescence.

Analysis of the *V. fischeri cyaA* locus

Analysis of the cloned *cyaA* locus of *V. fischeri* was performed at the sequence level and the construction of deletion mutants. We placed emphasis on the construction of a 'clean' deletion mutant, so that results obtained with these mutants are not confounded by the effects of heterologous DNA, which we have shown can alter the growth characteristics of *V. fischeri* and influence expression of the *lux* operon in artefactual ways. At the sequence level, we are working to define the end of the *cyaA* sequence, which was not retained in clones in our *V. fischeri* genomic library. We hypothesize that the terminal portion of the *CyaA* coding region contributes to the substantial instability of *V. fischeri cyaA* clones, and that as a consequence, clones containing the intact gene are not represented in the library. We have been informed by colleagues that the genomic sequence of *V. fischeri*, though not the strain under study in our laboratory, soon will be released, and information from that massive sequencing project may provide the sequence data we need for a more complete understanding of the *cyaA* gene. Following construction and physical verification of the appropriate *cyaA* mutations, these mutations will be delivered into MJ-100, MJ-215 *(*Δ*luxI, ainS*), and MJ-208 (Δ*luxR*) to construct *cyaA* mutants of these strains, which will then be analyzed for *Lux* expression.

Continuous-culture studies of *Lux* regulation.

Continuous cultures of *V. fischeri* MJ-100 under carbon-limitation have been performed. Work to this point has been directed at defining the appropriate growth medium and aeration conditions, and we are now assessing the effects of population density on induction and maintenance of induction of the *Lux* system. With these cells, we also are pursuing direct measurement of cAMP, methods for which we are developing using the pre-programmed Biotrak plate reader, provided by Amersham, and an associated EIA reagent kit.

Regulation of lux by newly identified regulatory proteins.

We have participated in a study of the roles of the regulatory proteins LuxO and LitR on *lux* operon expression in *V. fischeri* (and *Vibrio harveyi*). LuxO, first identified in *V. harveyi*, and LitR, a homolog of *V. harveyi* LuxR, were identified recently in *V. fischeri* and shown to control luminescence. Our studies show intriguing parallels in how *V. fischeri* and *V. harveyi* regulate light production, even though their luminescence systems and fundamental regulation differ markedly. These results establish the participation of two newly identified regulatory proteins in the control of luminescence in *V. fischeri*, and they validate our hypotheses, based on physiological analysis, that the current (now previous) model for lux regulation in *V. fischeri* lacks one or more regulatory elements. As such, these results guided our on-going analysis of global control of *lux* operon expression in *V. fischeri*.

1.1.2. Stringent response in *Escherichia coli* and *Mycobacterium tuberculosis*

Molecular biological investigation of the stringent response in *Mycobacterium tuberculosis*

The stringent response in a number of bacteria, including *E. coli* and *M. tuberculosis*, is mediated by ppGpp[1], whose synthesis is greatly increased following a decrease in the availability of specific amino-acids. The main effect of ppGpp is at the level of translation, where it inhibits some genes (all types of stable RNA as well as other genes related to the growth machinery, such as RNA polymerase), and enhances others (especially those related to amino-acid synthesis). The phenomenon has been well studied in *E. coli*, where the link between amino-acid availability and ppGpp synthesis has been long identified in the function of RelA and SpoT proteins, which respectively catalyze the synthesis and hydrolysis of ppGpp. It has been shown that the strength of RelA is greatly increased in the presence of stalled translational complexes, which occur when an uncharged tRNA is bound instead of a charged one.

The dual-function Rel_{Mtb} protein from *Mycobacterium tuberculosis* catalyzes both the synthesis and hydrolysis of ppGpp. In our previous work (Avarbock, D., Avarbock, A. and Rubin, H. (2000) Biochemistry 39, 11640) we presented evidence that the full-length 738 amino acid (82 kDa) Rel_{Mtb} protein might catalyze its two opposing reactions at two distinct active sites. In the present work, we purified and characterized fragments of the full-length Rel_{Mtb} protein and confirmed this hypothesis. A fragment containing amino acids 87-394 (35 kDa fragment) has only ppGpp synthesis activity and a fragment containing amino acids 1-181 (20 kDa fragment) has only ppGpp hydrolysis activity. We also purified a fragment containing amino acids 1-394 (45 kDa fragment) that possesses both synthesis and hydrolysis activities. Unlike wild type Rel_{Mtb}, the synthesis activity of the fragments is not

[1] We use ppGpp as a shorthand for two substances, *guanosine tetraphosphate* (ppGpp) and *guanosine pentaphosphate* (pppGpp).

enhanced by the previously described Rel_{Mtb} activating complex (RAC), and the hydrolysis activity of the fragments is not inhibited by this complex. The basal hydrolysis k_{cat}/K_m of the fragments is decreased approximately 55-fold compared to the basal hydrolysis activity of the wild type Rel_{Mtb}, whereas k_{cat}/K_m for synthesis only decreased approximately 4-fold. In addition, wild type Rel_{Mtb} exclusively forms trimers and removal of the C-terminus results in the isolation of monomers. Therefore, Rel_{Mtb} catalyzes two opposing reactions at distinct active sites, and the C-terminus is involved in multimerization and regulation of both synthesis and hydrolysis. A general mechanism for transcriptional regulation by environmental amino acid concentration is suggested that applies to eukaryotes and prokaryotes. This work is reported in [7].

To understand Rel_{Mtb} -dependent genes that might contribute to this phenotype we compared the global transcriptional response of the wild type and the ΔRel_{Mtb} strain under stringent conditions. These studies revealed that the Rel_{Mtb} regulon encompasses a generalized down-regulation of the transcriptional apparatus as well as specific alterations in putative virulence factors and metabolic enzymes. The regulon also includes some heat shock proteins and secreted antigens which may alter immune recognition of the recombinant organism. These studies establish that the Rel_{Mtb} regulon is critical for the successful establishment of persistent infection in mice and begin to define immunologic and enzymatic factors of possible relevance to latent infection in humans. We have identified approximately 80 genes regulated by the stringent response. This work will be submitted for publication in the near future.

Dynamical model for the stringent response and the role of ppGpp in growth rate regulation

We constructed an ordinary differential equation (ODE) model to include several mechanisms that have been related to the stringent response in the literature reconstructing the dynamical feedback loop that determines the ppGpp concentration in the system, determining the rate of ribosome and RNAP production, and ultimately the growth rate. The mechanism of ppGpp mediated growth control is similar in *E. coli* and *M. tuberculosis*. This allows using the former, which is much easier to study under standard laboratory conditions, as a model for the latter.

Transcription mechanism The key to differential regulation by ppGpp is the distribution of RNA polymerase between different genes. We define several groups of genes, each with its own kinetic constants. Along with free RNA polymerase, we define an intermediate state consisting of an RNA polymerase occupying a promoter site, followed by an elongation state where the promoter is released but the RNA polymerase is engaged. The initiation state can break up back into a free promoter and RNAP, or can move into the elongation state. The elongation state is characterized by a long lifetime (we are also investigating a more careful description of elongation, see below). This choice of model states can accommodate several mechanisms that are proposed to explain the differential regulation effects of ppGpp, based either on the lifetime of the initiation state, or on the elongation time.

Translational elongation drives feedback We also define a translational elongation state. It is very important to properly account for translational elongation, since the paused elongation complexes play a crucial role in the stringent response by differentially regulating the activity of the enzyme Rel which controls both the synthesis and decay of ppGpp. We estimate the concentration of these complexes directly from the total translation rate and the charged/uncharged tRNA concentration. This closes the ppGpp – transcription – translation – RAC – ppGpp control loop, allowing us to make steady-state predictions for ppGpp versus amino-acid levels.

The preliminary stringent response model was used in [4] to study the application of our hybrid systems reachability algorithm.

In the process of building a comprehensive model for the stringent response in *M. tuberculosis* and *E. coli* we have shown that the same level of differential regulation can be achieved by many reasonable sets of translational kinetic parameters. Separate sets of parameters compatible with each of the two of the leading theories on this mechanism are possible. While differential regulation only may not be able to distinguish between these two views, model predictions on the time evolution of the stringent response do have the potential of [in]validating such models.

Continued study of the stringent response model revealed the possibility of bistability in the way the growth rate is determined. This finding may have important applications in understanding bacterial persistence.

1.1.3. Lactose regulation system in *Escherichia coli*

We investigated the Lactose metabolism model using reachability analysis and the approximate reachability algorithm. We were able to show that the presence of a basal rate of mRNA production is necessary for ensuring the switching properties of the system. This result is intuitive but not easily proven mathematically using traditional methods.

We demonstrated how this result can be inferred using reachability analysis, based on a hybrid systems abstraction to a continuous ODE based model of the system.

We have also performed a sensitivity analysis of the original lactose metabolism model. We recalculated the steady states of the model for 'noisy' parameter sets obtained by adding controlled random variations to the original values. We found that (i) the bistability and switching properties of the model are conserved for parameter sets with up to 20% variations; (ii) our hybrid approximation to the original model leads to steady states that are within a variation range corresponding to 5% or less uncertainty in the parameter values; (iii) there is high sensitivity to selected parameters such as the basal transcription rate.

We extended our investigation of the lactose system to a model including the effects of glucose. Using the same type of approximations (ignoring the time delays; see discussion of piecewise linearization below) we applied to the lactose-only model, we were able to calculate the steady states of the extended system. The bistability seen previously is preserved. We calculated analytically the region in the lactose-glucose plane where the system has two stable steady states. Direct numerical simulations are consistent with bistability being associated with the above mentioned region.

Preliminary results on a piecewise linear approximation of the glucose-lactose model are encouraging in the sense that with reasonable approximations of the nonlinear rate functions of the model lead to a steady state structure that is qualitatively similar to the one found in the non-approximated model.

This extended model contains an example of a unique rate function describing the transcription rate of the *Lac* operon as a function of the concentrations of CAP and allolactose. The mathematical details of this mechanism are somewhat unusual, involving a partition function consisting of 50 algebraic terms,

and just as many parameters. This particular *analytical structure* is not likely to appear in any other model so it needs to be handled on a case-by-case basis, which poses an unexpected challenge for our software.

Finally, we performed reachability calculations on the Lactose-glucose model. We focused on the notion of 'inducibility', the possibility of causing the system to settle in a high Lactose-processing state, starting from an initial state basically devoid of both allolactose and the gene products of the *Lac* operon. Inducibility as defined here can readily be formulated as a reachability problem. We found very good agreement between the region of bistability and reachability results. We showed that a necessary condition for inducibility is that during its evolution, the system must be outside the region in the Lactose-glucose plane that corresponds to multiple steady states.

An interesting byproduct of the extensive direct simulations performed with the model is that the method of identifying the threshold Lactose concentration used by the authors of the model paper is unexpectedly inaccurate.

We also used the *Lac* model as a test bed for a number of alternative approaches.

One new approach identifies elliptical sets of points in state space that are inside the region of attraction using sum of squares decomposition of the exact equations of motion. For this purpose we devised a reduced, three-dimensional version of the *Lac* model. The initial results are encouraging; we were able to cover a significant fraction of the two regions of attraction with single ellipsoids and approximate the dividing surface. The results are published in [9].

An approximate reachability algorithm was initially formulated as a heuristic. We are currently investigating its relationship to notions of flow in state space and stochasticity. The latter connection is potentially very promising, as we may be able to frame our approach as a highly efficient method of simulating the stochastic master equation (which describes the time evolution of the *probability density* of the states of a stochastic system). This direction is still under study.

1.1.4. Tetracycline resistance in *Escherichia coli*

The wide use of tetracycline, a popular broad-spectrum antibiotic, has been accompanied by the spread of resistant bacteria. The basic mechanism of tetracycline resistance is associated with proteins encoded by genes located on moveable transposons or resistance factors which can be easily exchanged or distributed among bacteria, as a result of which antibiotic resistance spreads quickly.

An important tetracycline resistance is due to the efflux pump affected by the TetA protein which is transcribed by the *tetA* gene present in resistant bacteria. The TetA protein is itself toxic to the prokaryotic cells, and the mechanism provides a very tight but efficient switch for gene expression, which is sufficiently modular and self-sufficient. This mechanism has been adequately studied for workable mathematical models to be constructed. The goal of mathematical models is design of *in silico* experiments so that the dynamic reaction of the cells to tetracycline can be understood, which will have a definite impact on drug usage.

The mathematical model of the tetracycline display no positive feedback loop as a result of which there exists no automatically discernible multi-stable state. Our extensive search on parameter values, which we have constrained, using model and physical arguments, do not indicate that multi-stable states exist.

Simulation of the model from realistic initial values of the variables indicate that the internal tetracycline concentration spikes to a large value for periods longer than 1000 seconds, which is much longer than the lifetime of the bacteria. This indicates that the cell dies in such conditions purely due to the dynamic effect of the TetA protein that is transcribed by the *tetA* gene and which provides the pump that expels the drug from the cytoplasm into the periplasm The TetA transcription is slow and this allows the cytoplasmic tetracycline concentration to rise above 10 μM, thus inhibiting the ribosome, and leading finally to death of the cell. Thus the dynamic effects of TetA further constrain the possible parameters that describe the tetracycline resistance mechanism in *E.Coli.* This work will be submitted for publication in the near future [12].

1.2. Metabolic network modeling

Our approach to metabolic networks attempts to identify the constraints on the metabolic output of a network of reactions that follow from reaction stoichiometry. Our long term vision is to be able to *enumerate* the metabolic states that are compatible with a given set of exterior constrains, such as availability of nutrients.

1.2.1. Producibility

We developed an algorithm that constructs all feasible states that have a net production of one metabolite, given a set of available nutrients. Application of our analysis tools to constrained stoichiometric metabolic models built from the BioCyc database allows identification and enumeration of database inconsistencies which lead to model "leaks". These inconsistencies result from the use of generic species, errors in stoichiometry, and missing species in reactions. We have also applied a novel approach which combines analysis of a genome-scale metabolic model with *in vivo* survival data for gene delete on mutants, allowing identification of sets of metabolites that are essential and non-essential for *E. coli* survival.

Using our approach of determining feasibility of metabolite production for large metabolic networks, we have generated *in silico* gene to metabolite knockout maps for a variety of nutrient media settings for *E. coli* using a recently published genome-scale metabolic model. From analysis we are able to identify gene deletions with large metabolomic impact as well as fragile metabolites in the context of these nutrient media. Combining our results with *in vivo* survival data and applying a variant of an association-rules data mining approach, we were able to identify complex Boolean combinations of metabolites that appear necessary for survival.

Combining the SBML compatible version of the producibility tool with the model repository at SBML.org, we are able to characterize leaks over a hundred KEGG database derived models. We are also able to do the same for BioCyc derived models parsed using Harvard's Biocyc2SBML translator. Leaks in metabolic pathway databases arise from mistakes in reaction annotation, mostly due to errors in stoichiometry, missing species, and inconsistent use of generic species. The presence of leaks renders flux balance analysis of metabolic network models inaccurate, since they introduce spurious sources and sinks that implicitly serve to provide nutrients or consume species.

1.2.2. New media

We have employed the extreme semipositive conservation relations (ESCR) of the genome scale metabolic model *E. coli* iJR904 to generate novel nutrient media that render growth feasible.

Underpinning our method is a duality between ESCR and the production capabilities of a metabolic network, which allows minimal nutrient sets to be generated for an arbitrary species in the network through a simple traversal of ESCR. Computing the non-water-containing ESCR for the *E. coli* iJR904 genome scale metabolic model, we employ our algorithm to determine all 928 minimal aqueous nutrient media that are stoichiometrically compatible with biomass production. Each aqueous nutrient set generated by our analysis is minimal in the sense that any of its water-containing subsets fail to render biomass producible in a fully reversible network. Applying irreversibility constraints, we find 287 of these 928 nutrient sets to be thermodynamically feasible. We also find that an additional 365 of these nutrient sets are thermodynamically feasible in the presence of oxygen. Our results correspond to testable hypotheses of alternate growth media derived from analysis of the *E. coli* genome scale metabolic network.

2. Algorithm Development

2.1. Tools for dynamic system analysis

2.1.1. Reachability of multi-affine systems

In our work, approach genetic regulatory networks are modeled as switched systems with rectangular invariants and multi-affine dynamics. We are using results from hybrid systems theory to investigate reachability properties of the networks. The idea of reachability is to verifying the existence of systems trajectories that connect two regions of state space, without recourse to direct simulations. Our main focus is building tools that implement an efficient reachability algorithm developed for piecewise multi-affine hybrid systems.

We also have developed an approximate version of the reachability algorithm. It is based on the same framework as the original one (multi-affine equations and rectangular partition), but links between two adjacent blocks are considered active if the flux from the first box into the second exceeds a threshold. This threshold is set as a fraction of the total flux through a facet. We have made significant progress in defining efficient algorithms based on this notion that quickly identify candidates for stable steady states.

We have implemented as a proof of concept a projection algorithm that allows for producing light image files for bi-dimensional projections of high dimensional reachability graphs. This procedure is specific for hyper-rectangular partitions and will be implemented soon in the Charon graphical representation module.

Also as a proof of concept we are currently investigating the direct solution of the multi-affine model equations in search of steady states. This method is not expected to scale well but it seems feasible for mid-size biochemical networks (10 to 100 species).

As a proof of concept, we applied a sum-of-squares algorithm to a three-dimensional version of the Lactose-only model. We were able to identify elliptical subsets of the regions of attraction of the steady states, for various values of external Lactose.

2.1.2. Building hybrid system abstractions for dynamical systems

We developed a procedure to build piecewise linear approximations to chemical rate laws, allowing the automated conversion of models given in SBML to a hybrid form. This involves a library of pre-defined linear approximations for a set of common rate laws, and a generic mathematical procedure to be applied to non-standard reaction types. Our procedure is based on an annotation scheme that identifies the reaction types that occur in a given model. The hybridization procedure is predetermined based on the type of rate law.

Hybrid Systems Model Builder (**HSMB**) combines two important functions necessary in constructing a hybrid systems model of a biological system. The first function of HSMB is to input an SBML model of the system and perform piecewise-linear approximation of the rate laws in the reactions of the model, based on the annotations in the SBML file. Changes to the parameters of the rate law are introduced using the SBML mechanism of events. The output of this stage is an SBML model. The second function of HSMB is to transform an SBML model (either original of with the approximated rate laws) into the input of the hybrid systems analysis tool Charon. In the last quarter, the functionality of HSMB was substantially extended with the ability to process arbitrary rate laws. A new type of annotation for complex rate law functions was introduced. This annotation indicates to the HSMB that the rate law is not one of the standard rate laws with pre-defined approximation functions. Such a rate law is approximated using a user-supplied grid. The approximation fits the new functions to the rate law values at the grid points. The change required a substantial re-design of both the hybridization module and the Charon translation module of HSMB.

2.2. Randomization

For given dynamic systems, a randomized method can be a good vehicle for algorithm design and state space search. We have developed a randomized searching method to verify control algorithms or generate suitable test vectors. We are working on integrating this into Charon. This should be finished within a month.

2.3. Tools for metabolic network analysis

2.3.1. Producibility of individual metabolites via linear optimization

We have developed an approach for systematically evaluating the biosynthetic capabilities of a metabolic network given a media and genotype. This algorithm uses a set of linear feasibility checks on the stoichiometry matrix to generate a set of producible metabolites. Sets of producible metabolites can be compared between wild type and mutant networks to generate sets of metabolite knockouts. Combining these results with *in vivo* phenotype data allows identification of putative essential metabolites for a given phenotype.

We have implemented a variant of the *apriori* association rules data mining algorithm to identify complex relationships between *in silico* metabolite production and *in vivo* phenotypes. This approach allows automatic identification of Boolean combinations of metabolites whose knockout associates with the disappearance of a particular phenotype, such as survival. Results from this approach can be used to build testable hypotheses regarding what metabolic capabilities are essential for a phenotype.

2.3.2. Identification of extreme conservation laws

We have developed a novel approach that determines minimal nutrient sets that render an arbitrary metabolite producible in a fully reversible genome-scale metabolic network. This approach employs the extreme semipositive conservation relations (ESCR) of a metabolic network, exploiting the duality between conservation and metabolite producibility. Following generation of ESCR, we employ a depth first traversal of ESCR combined with pruning to enumerate minimal nutrient sets. Our approach allows results to be obtained even from partial computations of ESCR, which facilitates its scalability to large metabolic networks. Our results, though directly applicable only to a fully reversible network can be evaluated in the context of thermodynamic constraints using the optimization-based test for producibility.

3. Software Implementation

3.1. Charon

We first implemented our discrete reachability analysis algorithm in Charon. A more thorough description of this algorithm can be found in Calin Belta, Luc Habets, and Vijay Kumar, "Control of multi-affine systems on rectangles with applications to hybrid biomolecular networks", CDC2002, Las Vegas NV, 2002. This algorithm divides phase space into a rectangular partition and generates a directed graph describing which partition elements are reachable from which of their adjacent elements. Once the directed graph has been generated, standard graph analysis routines can be performed to give weak evidence of reachability or strong evidence of non-reachability. At present, this algorithm is implemented to handle multi-affine systems, although we can extend it to handle more complex dynamics.

Several significant improvements were implemented in Charon version 2.0. We no longer need to compile code on the fly. This means that Charon no longer depends on the full java developers' kit, but can run on the standard java distribution. This allows us to avoid problems we have experienced that relate to java's inability to reload new versions of class files on the fly. This should also result in better efficiency. Several changes to the way Charon stores data internally should result in faster simulation. Charon now also has better support for non-deterministic mode-switching. We have created an installer for the Charon/Bio-Sketch Pad (BSP) combination, using the common Install Shield package. This allows non-technical users to easily install and use Charon and BSP.

We have developed a translator from SBML to Charon that allows us import general models into BioCharon for simulation and analysis. Files that furthermore conform to the proposed SBML annotation scheme mentioned above can be converted into piecewise multi-affine hybrid models, which can be either output in SBML form or can be readily processed by Charon.

3.2. Hybrid System Model Builder

Hybrid System Model Builder is a tool that automates the process of constructing a piecewise affine hybrid approximation to an arbitrary biomolecular network model given in SBML format. The core procedure builds piecewise linear approximations to the nonlinear rate laws specified in the input SBML file.

These approximations may be constructed in two different ways. For rate laws that are identified as one of the predefined types handled by HSMB, there is a built-in predefined piecewise linearization procedure which has been constructed to provide the closest approximation with a fixed number of linear pieces. The program recognized these rate laws through the annotation scheme (see below) that supplements the SBML markup language.

For rate laws not found in the predefined list, we provided an on-the-fly approximation modality. This procedure constructs piecewise linear approximations based on values given in a sequence of values (or rectangular mesh, in case of more variables). The resulting functions are piecewise linear or multi-linear and continuous.

The outputs of HSMB can be SBML, but its main function is to provide input to the Charon suite.

3.3. HybridSBML

HybridSBML is a toolset written in C++ that provides utilities for performing simulation and reachability tests of SBML models. The primary goal of this package is to provide high performance tools that are based on standard packages. It depends on the GNU Scientific Library (GSL) and a computer algebra system called GiNaC. The adoption of these packages gives it flexibility and a performance boost. The utilities take as input SBML models. The SBML files are parsed using the standard libSBML libraries. The package follows the special Biocharon annotations in the SBML needed for special tasks which are parsed with Xerces. The utilities are command line based. The two utilities included in this toolset are:

sbmlsim: This utility performs a simulation of the model, given initial concentrations encoded in the SBML. It can further perform piece-wise linearization of the model and simulate the linearized model. The linearization is performed according to the BioCharon annotation extension scheme of SBML.

sbmlreach: This utility performs a reachability analysis of the SBML model. The linearization is performed automatically on the correctly annotated SBML file.

3.4. Annotation scheme for SBML

We developed an annotation scheme designed to identify the most commonly used rate laws as well as their relevant parameters. The annotation scheme together with the library of motifs that would result from building it up is also applicable to software performing stochastic simulations, and will probably form the basis of a BioSPICE-wide standard.

The team has coordinated the use case for hybrid systems analysis involving NYU, SRI, U. Tennessee and Keck. We have ensured the interoperability of the respective tools, involving many small issues regarding SBML and compatibility in general. We were able to match simulation results indicating that the SBML files encoding the continuous and hybrid versions of the *Lac* model are correct. The same SBML files could be successfully used by ESS to deconstruct the complex rate laws into mass-action components. The resulting stochastic results are also consistent. We were able to demonstrate temporal logic queries in Simpathica (from NYU) on traces produced by Charon. The reachability results from SAL (from SRI) are consistent with ours

3.5. Metabolic analysis tools

The **producibility tool** was integrated into the dashboard and made compatible with SBML. This tool computes the producible set of metabolites given an SBML metabolic network model, set of nutrients, and set of knocked out reactions. It consists of a nutrient and reaction selector for input, and MATLAB code which uses semidefinite programming to determine the producible metabolite set. The producibility tool is part of the April "Debugging the Bug" use case, coordinated by Harvard University.

MinMedia Given a set of ESCR corresponding to a genome scale metabolic model and an objective species, returns a collection of minimal nutrient sets that render that specie's metabolites producible.

3.6. Utilities

We developed the software utility **Decimator** as part of our multi-institution use case for the DARPA BioSpice program. It converts time series data to regular, user specified time samples. This allows us to use arbitrary simulation or experimental data results with NYU's Simpathica software.

4. Publications

1. Belta, C., Finin, P., Halasz, A., Imielinski, M., Kim, J. W., Kumar, V. and Rubin, H. "Modeling and analysis of metabolism and the stringent response in *Mycobacterium tuberculosis*," *DARPA BioComp proceedings*, May 2003.

2. Esposito, J., Kim, J. W. and Kumar, V. "Adaptive RRT's for Validating Hybrid Robotic Control Systems," *6th International Workshop on Algorithmic Foundations of Robotics*, Utrecht, Netherlands, July 11-13, 2004.

3. Rubin, H. "Bacterial Response Mechanisms: Nature's Own Hybrid System", invited talk at the *7th International Workshop on Hybrid Systems: Computation and Control*, Philadelphia, 2004.

4. Belta, C., Finin, P., Habets, L.C., Halász, Á., Imielinski, M., Kumar, V., and Rubin, H. "Understanding the Bacterial Stringent Response Using Reachability Analysis of Hybrid Systems", in *7th International Workshop on Hybrid Systems: Computation and Control. Lecture Notes in Computer Science 2993*. eds. Alur, R. and Pappas, G. J. 111-125, 2004.

5. Esposito, J., Kim, J. W. and Kumar, V. "Adaptive RRT's for Validating Hybrid Robotic Control Systems," *6th International Workshop on Algorithmic Foundations of Robotics*, Utrecht, Netherlands, July 11-13, 2004.

6. Imielinski, M., Belta, C., Halász, A., Rubin, H., "Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities," *Bioinformatics*. 2005 May 1; 21(9):2008-16. Epub 2005 Jan 25.

7. Avarbock, A., Avarbock, D., Teh, J.-S., Buckstein, M., Wang, Z.-M. and Rubin, H. "Functional Regulation of the Opposing (p)ppGpp Synthetase/Hydrolase Activities of Rel_{Mtb} from *Mycobacterium tuberculosis*", *Biochemistry* 44:9913-9932, (2005)

8. Imielinski, M., Belta, C., Rubin, H., Halász, A., "Systematic analysis of conservation relations in *E. coli* genome-scale metabolic network reveals novel growth media," accepted for publication, *Biophys J*.

9. Ahmadzadeh, A., Halász, A., Prajna, S., Jadbabaie, A. and Kumar, V. "Analysis of the Lactose metabolism in *E. coli* using sum-of-squares decomposition." submitted to *ECC-CDC 2005*, Seville, Spain

10. Halász, Á., Kumar, V., Imielinski, M., Belta, C., Sokolsky, O., Pathak, S., Rubin, H. "Analysis of lactose metabolism in *E. coli* using Reachability Analysis of Hybrid Systems", *in preparation*

11. Pathak, S, Halász, Á., Kumar, V. and Goullian, M. "Modeling the *TetA* gene triggered tetracycline resistance in *E. coli* and the effect of membrane permeation on bacterial fatality", *in preparation*

5. Personnel

1. Dr. Harvey Rubin, Professor, School of Medicine and School of Engineering and Applied Science.

2. Dr. Vijay Kumar, Professor, School of Engineering and Applied Science.

3. Dr. Oleg Sokolsky, Professor, School of Engineering and Applied Science.

4. Dr. Adam Halasz, Research Scientist, School of Engineering and Applied Science.

5. Peter Finin, Research Associate, School of Engineering and Applied Science.

6. Jongwoo Kim, Doctoral Student, School of Engineering and Applied Science.

7. Marcin Imielinski, Doctoral and MD student, School of Medicine.

8. Seth Berger, student, School of Arts and Sciences.

A new company, Bio Software Systems, Inc., was started to develop transition products from University of Pennsylvania technology.